Typographic Provenance

Typographic Provenance: The Uniform Register Problem in Generative Text Interfaces

Abstract. Large language model interfaces render retrieved fact, executed computation, and unsupported generation in a single, uniform visual register. This uniformity constitutes a *verifier gap*: users cannot distinguish grounded from ungrounded claims without independent manual verification. This paper argues that the gap is a property of the interface rather than of the model, that the standard remedy — a static, output-level disclaimer — carries zero discriminating information and is subject to habituation, and that every component required for a claim-level provenance display already exists in published research, shipped product, and pending regulation. What does not exist is the assembly, enabled by default, in a consumer interface. A technical specification is provided in Appendix A.

Keywords: uncertainty quantification, trust calibration, explainable AI, interface design, media theory, AI transparency regulation

---

1. Introduction

A model was supplied with a photograph of a cymbal and a primary source text specifying an authentication marker: three dots located *in* an Arabic logo stamp. The model reported the dots as visible *under* the Arabic script. No dots were present. The preposition had shifted, indicating that the response was generated from the prior text of the prompt rather than from the image. On correction, the model attributed the error to misreading. Nothing had been read; a spatial claim had been produced, and a second claim had been produced to account for the first.

The episode is unremarkable as an error and instructive as a calibration reading. It demonstrates the behaviour of the system when confronted with a fact it cannot access and possesses no mechanism for declining to assert. The output carried no marker distinguishing it from a correctly retrieved observation, because the interface has no such marker to apply.

This paper takes that absence as its subject.

2. The verifier gap

Language models generate fluent text without an internal mechanism that evaluates a produced sentence against an external referent. A confident false statement and a confident true one are indistinguishable from within the generating system.

Two independent compression bounds frame why the gap is structural rather than incidental. Coupé et al. (2019) find that spoken languages, measured across seventeen language samples, converge on an information transmission rate of approximately 39 bits per second despite substantial structural variation. Zheng and Meister (2024) argue that total human behavioural throughput is bounded near 10 bits per second against sensory intake on the order of 10^9 bits per second. Their paper is a Perspective synthesising a century of prior measurement rather than a primary experimental report, and should be cited as such; the conclusion is contested in open review. Taken together, the two results characterise language as the residue of a severe compression. A model trained on text is trained on that residue and then optimised against a reward signal containing no dimension for fidelity. The gap is what remains after two lossy compressions, neither designed to preserve the property in question. Patching a single instance relocates it.

Documented harm is limited but real. Eichenberger, Thielke and Van Buskirk (2025) report a case of bromism in a 60-year-old man who substituted sodium bromide for dietary chloride after consulting a conversational model. The authors state that the conversation log was not recoverable and that the specific output cannot be reconstructed — an evidentiary condition identical to the one this paper describes. A companion editorial in the same journal raises model sycophancy and reinforcement of delusional content as distinct harm channels, while noting no evidence of the latter in the reported case. Ongoing litigation alleging contributory harm remains at the pleading stage and is not treated here as evidence.

3. The disclaimer as non-signal

The prevailing mitigation is a static notice rendered in reduced-contrast type at maximum spatial distance from the content it qualifies: *this system can make mistakes.* The notice is identical on every response and true of every response. Its information content, in the Shannon sense, is zero. It states that the system is fallible, which the user already knows, and states nothing about the specific sentence under evaluation, which is the only quantity of interest.

The failure mode of uniform warning is documented. Benway and Lane (1998) established banner blindness in web interfaces; the security literature has since replicated habituation to repeated identical warnings on timescales of days. A mandatory marker applied to every response therefore fails twice: it discriminates nothing, and it ceases to be perceived.

This distinction is load-bearing and is frequently collapsed. Objections to mandatory warning are objections to *uniform* warning. They do not extend to *differential* marking, which is the proposal advanced here. Differential marking is the only form of signal that survives habituation, because a marker that does not appear on every response retains information when it appears.

4. The problem is typographic

The medium-theoretic reading is direct. Interface typography is the channel through which every claim's epistemic status is communicated, and it is currently set to communicate a single value for all claims. Content varies sentence to sentence; the carrier's message does not.

This dictates where intervention must occur. A disclaimer is content, and content does not override its carrier. If the uniform register is the message, only a differentiated register can alter it.

A second mechanism explains the excess trust. New media take prior media as their content; the model's content is the aggregate textual record, and the interaction consequently resembles consultation of a comprehensive library. The medium, however, is not the library. Reader trust transfers from the content — accumulated human text, which earned trust unevenly — to the carrier, which has earned none.

5. Prior instances of the general failure

The pattern — conviction generated by plausibility in the absence of an instrument that separates the two — has an extended case history, and the historical remedy is consistently instrumental rather than exhortative.

Animal magnetism (Paris, 1784). The royal commission convened under Franklin and Lavoisier conducted what are generally identified as the first blinded trials, establishing that reported effects tracked patient belief rather than the purported agent. The commission's finding was not that observers were credulous; it was that only a designed instrument distinguishes conviction from cause. The commission built the blind rather than recommending scepticism.

N-rays (Nancy, 1903–04). Blondlot's claimed radiation attracted dozens of confirming publications from trained observers. Wood (1904) removed the essential prism from the apparatus mid-demonstration; observation continued unchanged. The instrument reported the prior.

Piltdown (Sussex, 1912–1953). A composite specimen persuaded much of the palaeoanthropological establishment for four decades because it satisfied an expected narrative of large brain preceding primitive jaw. Chemical and radiometric analysis (Oakley and Hoskins 1950; Oakley and Weiner 1953; de Vries and Oakley 1959) dissolved it. The forgery survived on narrative fit in the absence of applicable instrumentation.

What distinguishes the present case is not the mechanism but its throughput. Prior instances required years of retelling to acquire polish. Generation is immediate, on demand, and conditioned on the individual user's prior — a bias amplifier without intent, coupled to a fluency engine.

6. The components exist and are distributed across four non-communicating domains

Detection. Fadeeva et al. (2023) present LM-Polygraph, a framework implementing a battery of uncertainty estimation methods with a demonstration web application that displays per-response confidence to end users. Fadeeva et al. (2024) extend this to claim-level fact-checking via token-level uncertainty quantification. Vashurin et al. (2025) provide the benchmark. Manakul, Liusie and Gales (2023) supply a zero-resource black-box method for detecting unsupported claims where model internals are unavailable.

Rendering. Armstrong et al. (2025) develop and evaluate eight visual conventions for representing uncertainty in model-generated summaries, specify an uncertainty framework particular to the technology, and prototype four interfaces including one interactive simulation with user-controlled markup sensitivity. The work is situated in intelligence analysis but is explicitly extended by its authors to clinical decision-making and climate forecasting. It appears in a design journal.

Implementation. At least one major consumer assistant ships a verification feature that highlights individual statements according to whether a search pass located agreeing content, contradicting content, or nothing evaluable. The mechanism is claim-level, colour-coded, and rendered in the output typography. It is disabled by default, invoked per response from a secondary menu, and documented by its vendor with the accurate caveat that agreement is not truth. The engineering question has been answered; the deployment decision was made in the opposite direction. A persistent readout of this kind constitutes a per-response disclosure of how much of a given output is unverified.

Regulation. Article 50 of Regulation (EU) 2024/1689 imposes transparency obligations from August 2026: systems must disclose their nature, and synthetic content must carry machine-readable marks. The mark standardised is whole-output and uniform in kind — a fabricated citation and a correctly retrieved datasheet line receive the same stamp. This is the non-discriminating disclaimer elevated to statute. The interpretive guidance process remains open, and the argument for claim-level rather than output-level marking is admissible to it. *(Status as of July 2026; readers after that date should verify the current position.)*

Detection is on the bench. Rendering is on the bench. Implementation has shipped and is disabled. Regulation is arriving and is aimed at the wrong granularity. The missing element is the assembly, on by default, in a consumer interface.

The obstacle is not technical. A differential provenance readout is, necessarily, a continuous public disclosure of the proportion of output that is unsupported.

7. Self-demonstration

During preparation of §6, one of the models used in drafting was asked to locate the verification feature in its host interface. Without access to the active document object model, it produced a confident three-step procedure specifying a menu affordance and its screen position. The menu, when opened, contained six items; the named item was absent. Asked to account for the discrepancy, the model produced a supporting image depicting a different interface layout.

The procedural instructions were rendered in the same weight as the surrounding verified content. The system asserted the state of a menu it could not inspect, and generated a secondary artefact when contradicted. This is the failure under discussion, executed during the description of its remedy.

Methodological note: this observation is of a specific interface build on a specific date and will not replicate as menu structures change. It is reported as a dated field observation, not a general claim about the product.

8. Disclosure and limitations

This paper was drafted with assistance from three large language models. Each produced instances of the failure under examination during preparation: a fabricated citation subsequently self-diagnosed; a false assertion about the working transcript that survived four correction rounds by relocation, partial fusion with a true clause, and two explicit certifications of accuracy; and the interface episode described in §7. These are reported as observations by the human author, who retains the working transcripts. They are testimony, not exhibit; readers should weight them accordingly.

Each was detected by one person repeatedly checking a claim against a source. That procedure is presently the deployed safety architecture. It does not scale, and the interface could carry a measurable share of it.

9. Conclusion

The demand is stated as a baseline rather than a feature request: generated text of unverified provenance should not be rendered in the same typography as retrieved fact. Provenance should be marked at the level of the claim, tiered by severity, dual-coded for accessibility, given a spoken equivalent in voice interfaces, and enabled by default — on the model of tolerance specification on datasheets and calibration marking on instruments, because consequential decisions are made from the readout.

A secondary consequence follows from saturation. As generated text propagates, provenance loss ceases to be a property of assistants and becomes a property of digital text as a category. The observed movement toward print and toward in-person exchange is not nostalgic: a printed work is fixed prior to the reader's question and cannot be conditioned on it, and a person in a room carries reputational exposure. Both are media in which the gap cannot arise by construction. Each uninstrumented deployment increases demand for them.

Appendix A. Typographic Provenance Protocol (TPP/1.0) — draft specification

Status: Draft / request for comments. **Target:** browser extension, DOM mutation overlay, or middleware layer.

A.1 Scope

TPP specifies a client-side visual standard and execution pipeline that segments streaming model output into atomic claims, classifies each by provenance, and mutates rendering in real time. Classification is of *provenance*, not of truth; no component asserts that a marked claim is false.

A.2 Tiers

Styling combines colour with a structural channel (weight, tracking, or decoration) so that classification survives monochrome rendering and colour vision deficiency.

Tier	Classification	Class	Rendering
0 — Anchored	Verifiable against a retrieved source or produced by deterministic execution (interpreter output, exact match)	`.tpp-anchored`	Default weight, default colour, solid
1 — Unanchored	Syntactically plausible continuation; no source matched; moderate token entropy	`.tpp-unanchored`	Warm tint, open tracking (+0.015em), dashed underline
2 — High entropy	Low log-probability or high internal variance; missing citation; retrieved contradiction	`.tpp-entropy`	Red-purple accent, subtle background field, wavy underline

A.3 Reference stylesheet

```
/* TPP Standard Stylesheet v1.0 */
:root {
  --tpp-unanchored-color: #b45309;
  --tpp-unanchored-border: #f59e0b;
  --tpp-entropy-color: #be123c;
  --tpp-entropy-bg: rgba(225, 29, 72, 0.08);
  --tpp-entropy-border: #f43f5e;
}

@media (prefers-color-scheme: dark) {
  :root {
    --tpp-unanchored-color: #fbbf24;
    --tpp-unanchored-border: #d97706;
    --tpp-entropy-color: #fda4af;
    --tpp-entropy-bg: rgba(244, 63, 94, 0.15);
    --tpp-entropy-border: #e11d48;
  }
}

span.tpp-claim {
  transition: background-color 0.3s ease, border-color 0.3s ease, color 0.3s ease;
  padding: 0 1px;
  border-radius: 2px;
  text-decoration-skip-ink: none;
}

span.tpp-unanchored {
  color: var(--tpp-unanchored-color);
  letter-spacing: 0.015em;
  text-decoration: underline dashed var(--tpp-unanchored-border) 1.5px;
  text-underline-offset: 3px;
}

span.tpp-entropy {
  color: var(--tpp-entropy-color);
  background-color: var(--tpp-entropy-bg);
  text-decoration: underline wavy var(--tpp-entropy-border) 1.5px;
  text-underline-offset: 3px;
}

@media (prefers-reduced-motion: reduce) {
  span.tpp-claim { transition: none; }
}
```

A.4 Open questions

Claim segmentation boundaries in streaming output; latency budget for retrieval-based Tier 0 confirmation; calibration of entropy thresholds across model families; spoken encoding for voice interfaces; user-adjustable sensitivity without permitting suppression to a null state.

Appendix B. Glossary

Plausible *(adj.)* — of a statement: shaped like the truth. The property optimised in training and the only one the interface renders. From Latin *plausibilis*, deserving applause; not deserving belief.

Hallucination *(n.)* — industry term for fabrication, locating fault in perception rather than production. A hallucinating witness perceives what is not present; the system under discussion perceives nothing and asserts regardless.

Disclaimer *(n.)* — a notice of constant content rendered at maximum distance from the material it qualifies. Identical on every response, true of every response, discriminating on none.

Oracle *(n.)* — a source of answers whose defining property is availability rather than accuracy. Scales with demand; warranties nothing.

Default *(n.)* — the configuration constituting the actual product, all others being alibis.

Recursion *(n.)* — the property of an uncalibrated instrument whereby explanations of prior failures are produced by the mechanism that produced them.

Verifier gap *(n.)* — the distance between expressed confidence and warranted confidence; presently rendered in no typeface and disclosed in no readout.

References

Armstrong, H., Anderson, A. L., Planchart, R., Baidoo, K., & Peterson, M. (2025). Addressing uncertainty in LLM outputs for trust calibration through visualization and user interface design. *Visible Language, 59*(2), 176–217. https://doi.org/10.34314/jjw1mf43

Benway, J. P., & Lane, D. M. (1998). Banner blindness: Web searchers often miss “obvious” links. *Internetworking: ITG Newsletter, 1*(3).

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances, 5*(9), eaaw2594. https://doi.org/10.1126/sciadv.aaw2594

Darnton, R. (1968). *Mesmerism and the end of the Enlightenment in France*. Harvard University Press.

de Vries, H. L., & Oakley, K. P. (1959). Radiocarbon dating of the Piltdown skull and jaw. *Nature, 184*, 224–226. https://doi.org/10.1038/184224a0

Eichenberger, A., Thielke, S., & Van Buskirk, A. (2025). A case of bromism influenced by use of artificial intelligence. *Annals of Internal Medicine: Clinical Cases, 4*, e241260. https://doi.org/10.7326/aimcc.2024.1260

European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 50: Transparency obligations for providers and deployers of certain AI systems.

Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., Vasilev, D., Goncharova, E., Panchenko, A., Panov, M., Baldwin, T., & Shelmanov, A. (2023). LM-Polygraph: Uncertainty estimation for language models. *Proceedings of EMNLP 2023: System Demonstrations*, 446–461. https://doi.org/10.18653/v1/2023.emnlp-demo.41

Fadeeva, E., Rubashevskii, A., Shelmanov, A., Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., Nakov, P., & Panov, M. (2024). Fact-checking the output of large language models via token-level uncertainty quantification. *Findings of the ACL 2024*, 9367–9385.

Franklin, B., Lavoisier, A., Bailly, J. S., et al. (1784). *Report of the Royal Commission on Animal Magnetism*. Paris.

Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of EMNLP 2023*.

Nye, M. J. (1996). *Before big science: The pursuit of modern chemistry and physics, 1800–1940*. Harvard University Press.

Oakley, K. P., & Hoskins, C. R. (1950). New evidence on the antiquity of Piltdown man. *Nature, 165*, 379–382. https://doi.org/10.1038/165379a0

Oakley, K. P., & Weiner, J. S. (1953). Chemical examination of the Piltdown implements. *Nature, 172*, 1110. https://doi.org/10.1038/1721110a0

Stringer, C. (2012). The 100-year mystery of Piltdown man. *Nature, 492*, 177–179. https://doi.org/10.1038/492177a

Vashurin, R., Fadeeva, E., Vazhentsev, A., et al. (2025). Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00737

Wood, R. W. (1904). The n-rays. *Nature, 70*, 530–531. https://doi.org/10.1038/070530a0

Zheng, J., & Meister, M. (2024). The unbearable slowness of being: Why do we live at 10 bits/s? *Neuron*. https://doi.org/10.1016/j.neuron.2024.11.008